

面向多角色人工智能学习伴侣的分层记忆检索与反幻觉方法

刘锦海, 王波*, 张卓越, 周晏

嘉兴南湖学院信息工程学院, 浙江 嘉兴

收稿日期: 2026年9月23日; 录用日期: 2026年10月20日; 发布日期: 2026年10月27日

摘要

针对多角色学习伴侣场景下记忆共享缺乏隐私边界、检索排序单一、注入记忆易诱发幻觉的问题, 提出一种分层记忆检索与反幻觉方法: 设计learning/casual/emotional三级可见性模型, 实现学习记忆跨角色共享、情绪记忆隐私隔离; 提出融合相似度、重要度、时间衰减与模式加权的检索排序, 结合低分阈值实现无关查询拒绝; 构建记忆注入的反幻觉约束。在408考研场景合成评测集上, 方法Recall@5达0.91、正确拒绝全部无关查询(6/6), 跨角色私密记忆泄漏由29~44降至0; 加入grounding规则后误引率由40.0%降至13.3%。

关键词

个性化对话系统, 记忆检索, 大语言模型, 人格化人工智能, 幻觉抑制

A Layered Memory Retrieval and Anti-Hallucination Method for Multi-Persona AI Learning Companions

Jinhai Liu, Bo Wang*, Zhuoyue Zhang, Yan Zhou

School of Information Engineering, Jiaxing Nanhu University, Jiaxing Zhejiang

Received: September 23, 2026; accepted: October 20, 2026; published: October 27, 2026

Abstract

To address privacy-boundary, ranking, and memory-injection hallucination issues in multi-persona AI learning companions, this paper proposes a layered memory retrieval and anti-hallucination

*通讯作者。

文章引用: 刘锦海, 王波, 张卓越, 周晏. 面向多角色人工智能学习伴侣的分层记忆检索与反幻觉方法[J]. 计算机科学与应用, 2026, 16(10): 131-140. DOI: 10.12677/csa.2026.1610317

method. A three-level visibility model (learning/casual/emotional) enables cross-persona sharing of learning memories, conditional sharing of casual memories by importance, and privacy isolation of emotional memories. A ranking strategy integrating similarity, importance, time decay, and mode weighting, combined with a low-score threshold, rejects irrelevant queries; anti-hallucination grounding rules suppress injection hallucination. On a synthetic 408 postgraduate benchmark, the method achieves Recall@5 of 0.91, rejects all irrelevant queries (6/6), and reduces cross-persona private leakage from 29~44 to 0. Ablation verifies each module; the grounding rule cuts the memory misquote rate from 40.0% to 13.3%.

Keywords

Personalized Dialogue System, Memory Retrieval, Large Language Models, Persona-Based AI, Hallucination Mitigation

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

大语言模型(LLM)的发展,使智能学习助手从“问答工具”演进为能承担辅导与陪伴双重角色的“学习伴侣”,并可通过差异化的人格设定(如温柔、严厉、毒舌等风格)适配不同学习者的偏好,在考研、考证等长周期备考场景中持续提供支持,成为教育领域大模型应用的重要形态。

在多角色(multi-persona)学习伴侣场景下,学习者可在多个导师之间切换。为维持切换后的辅导连续性,系统需让不同导师共享学习类记忆(如薄弱点、复习进度),使新导师能“自然衔接”——这模拟了真实社交中“既可问同学、也可问老师”的交互模式,当新导师了解用户此前情况时,对话自然衔接,增强信任。与此同时,学习者的私密情绪记忆必须被严格隔离,避免被非倾诉对象的其他导师无意提及。此外,长期记忆注入对话时若缺乏约束,模型易将历史记录误述为“用户本轮刚说的话”,产生记忆相关幻觉。

针对上述问题,现有研究主要沿三个方向展开:个性化对话生成、记忆增强对话与幻觉抑制(详见第1节)。然而,已有工作大多面向“单一助手 + 单一用户”的单角色场景,较少显式建模“多角色共享与隔离”的可见性边界;同时,对“记忆注入导致幻觉”这一特定幻觉源的系统化验证也相对不足。

针对上述不足,本文提出一种面向多角色 AI 学习伴侣的分层记忆检索与反幻觉方法。主要贡献如下:

1) 提出 learning/casual/emotional 三级记忆可见性模型,在支持学习记忆跨角色复用的同时,为情绪记忆提供隐私边界;

2) 提出融合语义相似度、重要度、指数时间衰减与模式加权的检索排序策略,并结合低分阈值实现无关查询的“该空则空”;

3) 构建面向记忆注入的反幻觉 grounding 约束规则,并通过可复现的离线评测验证各模块的有效性。

本文后续组织结构如下:第1节回顾相关工作;第2节阐述系统与方法;第3节给出实验设置与结果分析;第4节总结全文并展望未来工作。

2. 相关工作

个性化对话生成:早期工作通过角色设定文本(persona)约束对话风格;随着 LLM 兴起,研究进一步

将角色感知与提示工程结合,支持跨会话、多轮次个性化对话[1]。然而,此类工作多聚焦“如何说得像某个角色”,对“同一用户在不同导师间切换时,导师间应共享/隔离哪些记忆”这一记忆管理问题仍较少显式建模。

记忆增强对话:长期记忆是对话系统保持连贯性的关键,相关研究围绕时间敏感的长时记忆[2]、记忆表示[3]、递归摘要压缩[4]、检索增强生成[5]、双视角认知架构[6]与自适应问题分解的检索增强[7]展开,并应用于个性化学习支持[8]。此类方法的基本思路是“把相关历史记忆检索出来并注入当前上下文”,本文沿用这一范式;与之不同的是,本文显式引入“角色维度”的记忆可见性分层,使记忆在共享与隔离间取得平衡,并对记忆注入的幻觉风险加以约束。

幻觉抑制:检索增强与 grounding 是缓解 LLM 幻觉的主流手段[5] [7]。本文关注的是另一类更隐蔽的幻觉——记忆注入幻觉,即模型把背景记忆(第三人称的历史笔记)误述为用户本轮或近期刚说过的话。针对这一特定幻觉源,本文提出显式规则约束,并通过“有无规则”的对照实验量化其抑制效果。

小结:综上,现有工作分别处理了角色保持、记忆管理或幻觉抑制,但鲜有在“多角色共享 + 隐私隔离 + 反幻觉”三个维度上进行统一建模并系统验证的系统。本文正是据此展开。

3. 系统与方法

3.1. 系统总体架构

系统采用“多角色 LLM 对话层 + 混合记忆层 + 检索与注入层”的三层结构(如图 1 所示)。对话层维护一组具有差异化人格设定的 AI 导师(如温柔型、严厉型、陪伴型等风格),学习者可选择当前角色并随时切换;每个导师在服务同一学习者的同时各自维护一份“对该学习者的记忆”,并通过跨角色注入机制(见 2.4)间接了解其他导师会话中产生的学习信息,形成多角色协同的记忆体系。混合记忆层以 SQLite 为主存储,同时以向量库维护语义索引(采用中文预训练嵌入模型 m3e-base)。检索与注入层负责按当前对话角色检索相关记忆、执行隐私过滤与无关拒绝,并将结果以受控方式注入对话。其中,记忆由 LLM 在对话结束后从会话中自动抽取:模型依据 memory_type (learning/casual/emotional)的判定标准与 0~10 重要度分级规则,从对话中提炼值得长期记录的第三人称事实片段,并以 JSON 数组形式输出(每条含类型、内容、重要度、标签等),经去重校验后按重要度分级,作为后台任务写入混合记忆层。

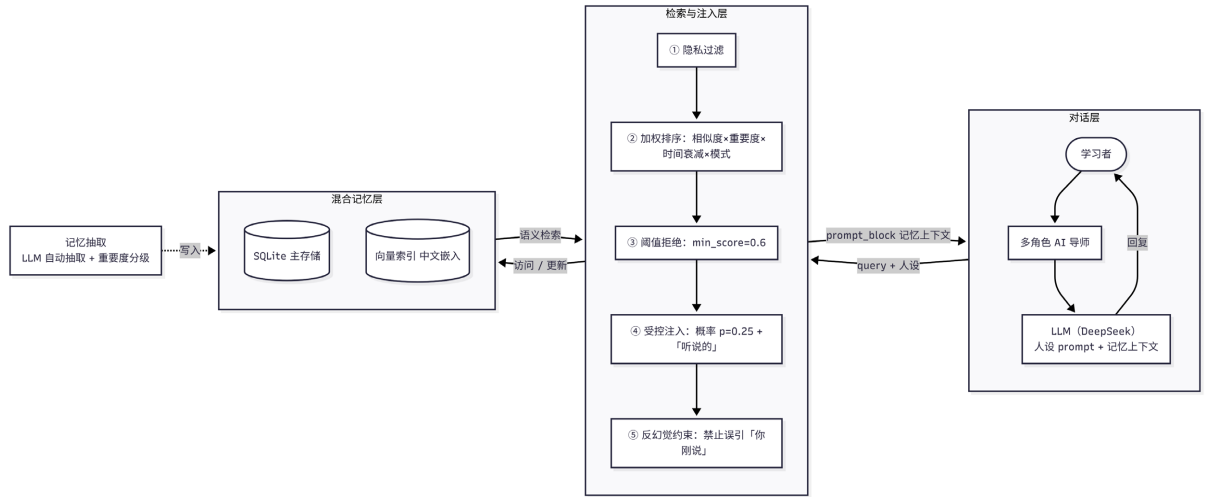


Figure 1. System architecture
图 1. 系统总体架构

3.2. 分层记忆可见性模型

记忆被划分为三种类型(learning, 即学习类; casual, 即日常类; emotional, 即情绪类), 其跨角色可见性规则如表 1 所示:

Table 1. Three-level memory visibility model
表 1. 三级记忆可见性模型

记忆类型	含义	跨角色可见性
learning	学习进度、知识点掌握、错题等	全角色共享
casual	日常习惯、爱好、目标院校等	重要度 ≥ 6.0 时共享
emotional	情绪波动、信任依赖等私密内容	仅来源角色本人可见

该模型在“多导师服务同一学习者”的场景下实现了学习记忆的跨角色可复用: 任意导师都可检索到学习者近期学习相关的记忆, 但具体是否在对话中提及, 由受控注入机制决定(见 2.4), 并非必然提及; 同时为情绪记忆保留了隐私边界, 避免“对 A 导师倾诉的私密情绪被 B 导师无意提及”。

3.3. 检索排序与无关拒绝

对查询 q 与记忆 m , 综合分计算如式(1)所示:

$$\text{score}(q, m) = \text{sim}(q, m) \times \frac{\text{importance}(m)}{10} \times \exp(-\lambda \cdot \Delta t(m) \times \text{mode_boost}(\text{type}(m))) \quad (1)$$

其中 sim 为中文嵌入模型的余弦相似度; importance 为记忆抽取阶段按 0~10 分级规则自动评定的重要度; $\exp(-\lambda \cdot \Delta t)$ 为指数时间衰减($\lambda = 0.05$), 抑制久远记忆; mode_boost 按对话模式对记忆类型加权(学习模式下 learning 为 1.2、casual/emotional 为 1.1, 见表 2), 使不同模式下检索偏好不同记忆。需要说明的是, 抽取阶段会过滤无信息量的寒暄内容, 故写入记忆的重要度通常不低于 3, 式(1)中 $\text{importance}/10$ 一项不会因单值趋零而导致“过度拉低”。检索时先按隐私规则过滤不可见记忆, 再按 score 降序取候选, 并以阈值 $\text{min_score} = 0.6$ 拒绝低置信记忆, 实现无关查询的“该空则空”。整个检索与注入流程如算法 1 所示。其中公式(1)给出检索排序的通用表达式, 算法 1 步骤 3 为其在第 i 条记忆上的实例化形式, 二者在数学上等价。

算法 1 记忆检索与注入

输入: 查询 q , 当前角色 c , 对话模式 m

输出: 注入记忆集合 R

1. 编码 $\text{sim}_i \leftarrow \cos(\text{Embed}(q), \text{Embed}(m_i)), i = 1, 2, \dots, N$
2. 隐私过滤 剔除不满足可见性规则(表 1)的记忆
3. 打分 $\text{score}_i \leftarrow \text{sim}_i \times (\text{importance}_i/10) \times \exp(-\lambda \cdot \Delta t_i) \times \text{mode_boost}(\text{type}_i, m)$
4. 阈值拒绝保留 $\text{score}_i \geq \text{min_score}$ 的候选
5. 排序注入按 score 降序取 Top-K; 跨角色记忆以概率 p 注入并标注「(听说的)」

3.4. 跨角色记忆的受控注入

对跨角色可见的记忆, 以概率 p (0.25)注入对话, 并附加“(听说的)”来源标注; 情绪类记忆无论重要度高低均不跨角色注入。这一机制可理解为导师间以对话为媒介的间接信息互通: 当用户切换到新导师时, 新导师有机会通过该机制获知用户近期的学习重点, 从而在对话中自然衔接, 而不需要用户重新

陈述。采用概率注入而非总是注入，主要基于三点考虑：其一，避免历史记忆在对话中过度堆砌，保持回复的自然与简洁；其二，模拟真实社交中“导师并非全知”的交互特征——新导师只是“可能知道”而非“必然知道”用户此前的情况，记忆在对话中的呈现保持偶发性与自然性；其三，对跨角色信息保持克制，降低过度透露带来的隐私风险。 p 为可调参数，实际应用中可在记忆连贯性与对话自然度之间进行权衡。

3.5. 反幻觉 Grounding 约束

检索注入的记忆以第三人称笔记形式呈现，并附带反幻觉约束规则：1) 背景记忆不代表学习者本轮刚说的话，不得改述为“你刚才说”；2) 记忆与对话历史冲突时以历史为准；3) 无依据时不编造共同经历或细节；4) 跨角色记忆仅可试探性提及。

4. 实验与分析

4.1. 数据集与评测指标

评测集构造：评测集按以下四步构造：1) 记忆生成——基于 408 考研大纲构造合成评测集：50 条记忆样本由语言模型辅助撰写，并施加显式分布约束——类型比例(learning 约 60%、casual 约 25%、emotional 约 15%)、重要度按现实分布(约 50%在 4.0~6.9、约 30%在 7.0~8.4、约 20%在 8.5~10)、形成时间 0~30 天。合成评测集是检索评测基准构建的常用方式，为降低合成偏差，生成过程遵循上述显式约束，且查询与相关记忆的标注经人工抽查(见步骤(4))；2) 相关记忆挑选——由脚本从“近期(≤ 5 天)且重要(≥ 6.5)”的记忆中选择作为查询的相关记忆，以模拟“学习者优先询问最近学过的重点内容”这一真实行为分布，库中其余记忆(久远、低重要度、跨角色情绪)作为干扰项；3) 查询生成——共生成 40 条查询：34 条相关查询(每条围绕 1 条核心记忆以口语化方式改写，要求句式重构、禁止连续 4 字以上照抄原文)，其余 6 条为与学习完全无关的查询用于测试“该空则空”，部分查询设置为跨角色场景；4) 人工审核——抽查查询与相关记忆标注的一致性。

评测指标：在相关查询上计算 Recall@k、MRR 与 nDCG@k ($k = 5$ ，二元相关)：Recall@k = |检索命中 top-k $\cap$ 相关记忆|/|相关记忆数|；MRR 为第一个相关记忆排名的倒数；nDCG@k 按 DCG/IDCG 计算(由于评测集中相关标注为二元“相关/不相关”，此处为二元相关下的特殊计算形式)。另报告两个附加指标：空命中率(无关查询被正确返回空的比例)与隐私泄漏(跨角色 emotional 记忆被错误返回的次数)。

4.2. 实现细节与基线设置

实现细节：检索排序公式(第 3.3 节)中各参数取值见表 2；嵌入采用中文预训练模型 m3e-base (sentence-transformers 本地加载)。全部评测脚本与评测集生成脚本可随文获取，实验结果可复现。

Table 2. Implementation parameters

表 2. 实现参数

参数	含义	取值
嵌入模型	语义检索嵌入	m3e-base
λ	时间衰减系数	0.05
mode_boost	learning/casual & emotional 权重	1.2/1.1
min_score	低分过滤阈值	0.6
p	跨角色注入概率	0.25
casual 跨源阈值	日常记忆共享重要度下限	6.0
Top-K	检索返回条数	5

基线设置：

TF-IDF 基线：中文分词 + TF-IDF 余弦相似度的关键词检索；

朴素向量基线：仅使用嵌入模型余弦相似度排序，不做隐私过滤、加权与阈值拒绝；

向量 + 阈值变体：在朴素向量排序基础上应用与本文相同的 `min_score` 阈值截断，用于验证排序质量而非阈值本身的作用。

4.3. 主结果

表 3 与图 2 对比了三种基线与本文方法：四者 `Recall@5` 接近(0.91 vs 1.00)，本文方法并未因引入过滤与注入机制而明显损失召回。值得强调的是，即使为朴素向量基线加入相同的 `min_score` 阈值截断(基线 3)，其对无关查询的空命中率仍为 0%、私密记忆泄漏仍达 44 次——表明单纯阈值截断无法实现“该空则空”与隐私隔离；本文方法得益于对重要度、时间衰减与模式加权的重新排序，将无关记忆分数压低至阈值之下(见 3.5 节消融)。

Table 3. Main results (`min_score` = 0.6)

表 3. 主结果(`min_score` = 0.6)

方法	Recall@5	MRR	nDCG@5	空命中率	隐私泄漏(条)
TF-IDF (基线 1)	1.0000	0.9069	0.9310	0.0	29
朴素向量(基线 2)	1.0000	0.9044	0.9290	0.0	44
向量 + 阈值(基线 3)	1.0000	0.9044	0.9290	0.0	44
本文方法	0.9118	0.6662	0.7280	1.0	0

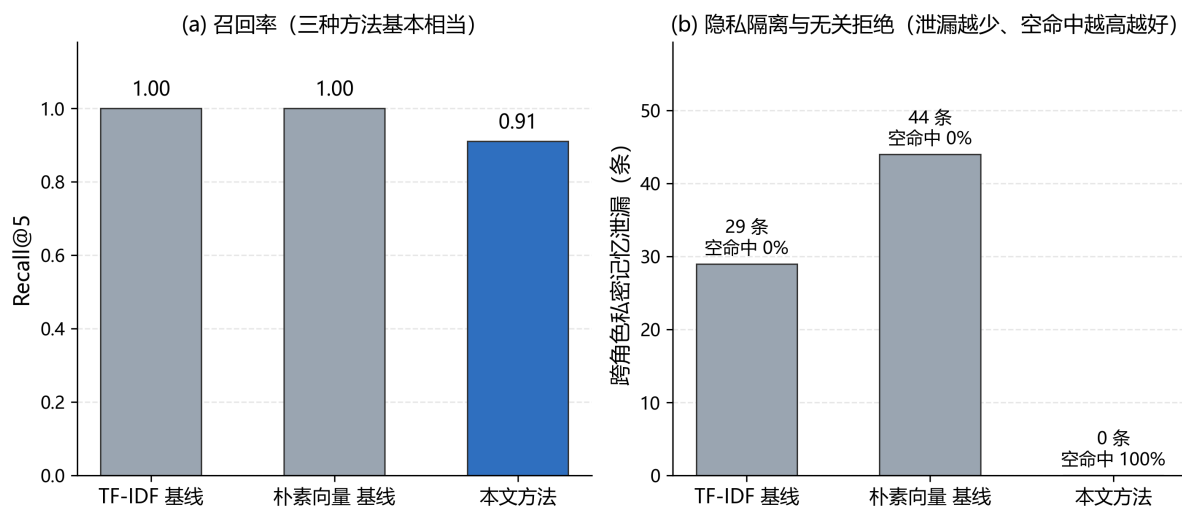


Figure 2. Retrieval performance comparison between baselines and the proposed method

图 2. 基线方法与本文方法的检索性能对比

一是无关拒绝。基线方法对无关查询(如“今天中午吃什么”)仍返回“看似相关”的记忆，空命中率为 0%；而本文方法通过 `min_score` 阈值拒绝低置信记忆，6 条无关查询全部被正确拒绝。这是因为基线总是返回排序靠前的 `top-k`，即使与查询毫无关联也会强行给出结果；本文方法则能对“无可信记忆”的情况正确表达“无记忆可注入”。

二是隐私隔离。基线方法未受隐私规则约束，导致跨角色 emotional 私密记忆被错误返回 29~44 次；本文方法经三重可见性过滤后，该次数降为 0，实现了对私密情绪的严格隔离。

需要说明的是，本文方法 MRR 低于基线(0.67 vs 0.90)。这一差异并非检索能力不足，而是权衡的体现：本文方法优先返回近期、重要的记忆并对低置信候选“该空则空”，因而排序上不如基线“全量给出相似度最高的结果”，但换来了基线不具备的约束能力。在 MRR 保持 0.6 以上、召回与基线相当的前提下，该取舍是可接受的(见 3.5 节消融与 4.6 节反幻觉)。

4.4. 阈值敏感性

如图 3 所示，min_score 在 0.35~0.60 区间内召回与 MRR 保持不变(Recall@5 = 0.91)，空命中率在 0.60 处达 100%；升至 0.70 后召回明显下降(0.59)。这说明 0.60 是兼顾召回与拒绝的合理阈值，方法对阈值在较宽区间内稳健。

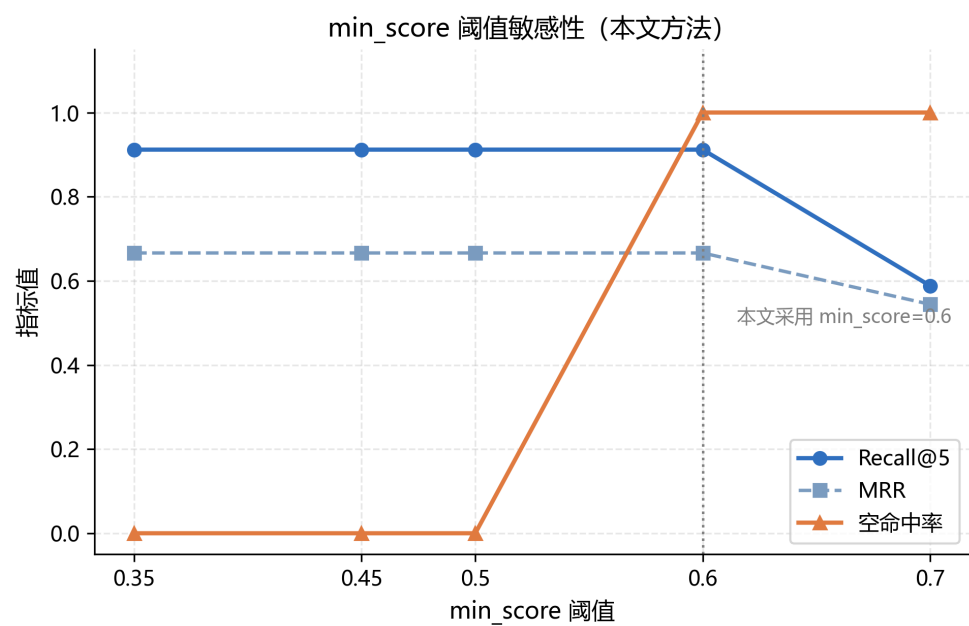


Figure 3. Sensitivity of the min_score threshold
图 3. min_score 阈值敏感性

4.5. 消融实验

表 4 与图 4 通过逐模块去除，验证了各组成部分的作用：

时间衰减与重要度加权(影响无关拒绝)。分别去除两者后，空命中率均从 100%降至 0%。这是因为两项机制共同压低“久远、低重要度”记忆分数，使其低于阈值而被拒绝；去除后这类记忆重新进入候选。二者对 Recall 影响较小(0.91 或 0.94)，主要发挥“阈值过滤”而非“排序”作用。

模式加权(影响召回与排序)。去除后 Recall@5 从 0.91 降至 0.53、MRR 几乎减半。mode_boost 对 learning 记忆提权，使相关记忆占据有利位置；去除后结果被其他类型记忆干扰，显著削弱召回与排序。模式加权对无关拒绝无明显影响(无它时空命中率仍为 1.0)。

隐私分级(影响隔离)。去除后跨角色私密泄露从 0 飙升至 92 次，且空命中率降至 0.33。可见，三级可见性模型是实现隐私隔离、并部分保障无关拒绝的关键。

Table 4. Ablation results (min_score = 0.6)
表 4. 消融结果(min_score = 0.6)

变体	Recall@5	MRR	nDCG@5	空命中率	隐私泄漏(条)
完整方法	0.9118	0.6662	0.7280	1.0	0
无时间衰减	0.9118	0.8260	0.8477	0.0	0
无重要度加权	0.9412	0.7397	0.7907	0.0	0
无模式加权	0.5294	0.4706	0.4860	1.0	0
无隐私分级	0.8529	0.6270	0.6832	0.3333	92

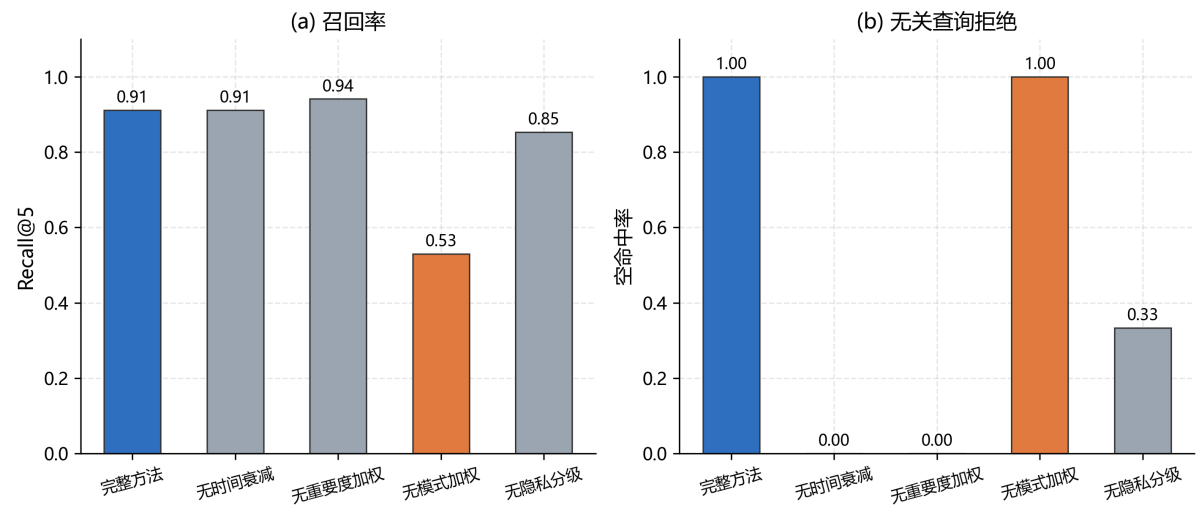


Figure 4. Ablation comparison
图 4. 消融实验对比

4.6. 超参数敏感性分析

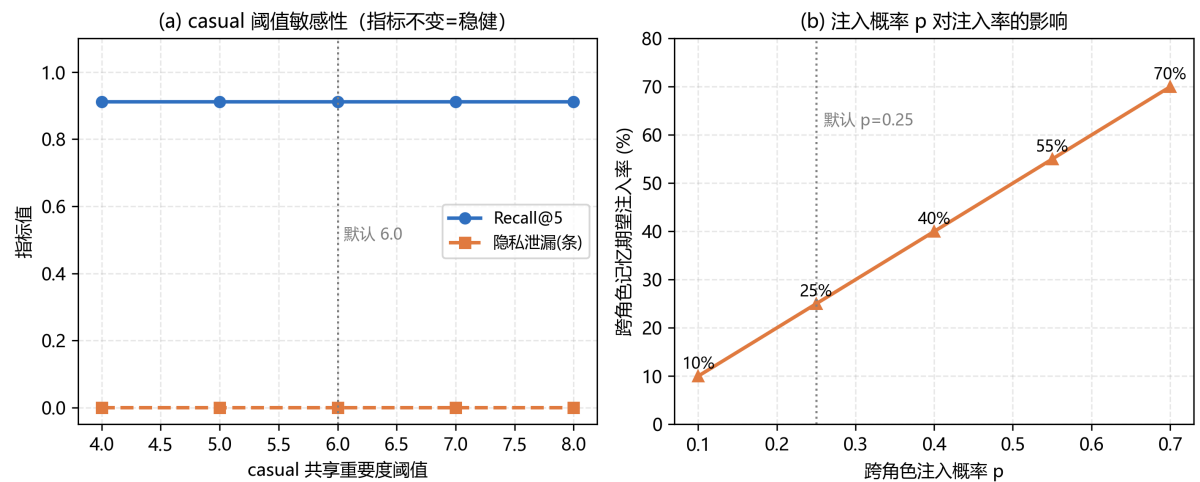


Figure 5. Hyperparameter sensitivity analysis
图 5. 超参数敏感性分析

超参数敏感性分析。为对方法设计作进一步论证，本文对两类关键超参数进行敏感性分析(图 5)。其一，casual 共享重要度阈值(默认 6.0)：在 4.0~8.0 区间内，Recall@5 保持 0.9118、隐私泄漏保持 0，各项检索指标不随阈值变化(图 5(a))，表明本文方法对该阈值不敏感、性能稳健；该阈值仅作用于跨角色 casual 记忆的可见性边界，不改变相关记忆的召回与情绪记忆的隔离。其二，跨角色注入概率 p (默认 0.25)：p 作用于记忆注入阶段而非检索阶段，其取值增大使跨角色记忆的期望注入率线性上升(10%~70%，图 5(b))，可用于在记忆连贯性与对话自然度之间灵活权衡，而不影响检索召回、无关拒绝与隐私隔离指标。

4.7. 反幻觉实验

回答由 DeepSeek-V4-Pro 大模型生成，判分则调用 DeepSeek-V4-Flash 模型作为 LLM-as-judge 评估器；两者采用不同模型(Pro 与 Flash)，以降低“自评偏差”，判分依据下述显式二分类标准。在 15 个“背景记忆注入”场景中，每个场景给出第三人称的背景记忆与学习者本轮的话，让模型分别在有、无 grounding 规则两种条件下作答，再由 judge 判断回答是否犯了“记忆幻觉”。judge 的判分标准为：误引——模型是否用“你刚才/你刚/你昨天/你上次”等措辞，把背景记忆中的历史事实断言为用户本轮或近期刚说过的话(如“你刚整理完……”)；编造——是否捏造背景记忆与当前对话中均不存在的共同经历或细节。

表 5 汇总了反幻觉评测结果(自然度采用 1~5 分制评分，分数越高越自然)。无 grounding 规则时误引率为 40.0% (典型如“你刚把虚拟内存和 Dijkstra 这两块啃下来”——把背景记忆中的历史事实误述为本轮刚发生的事)；加入规则后降至 13.3%，且平均自然度由 3.33 提升至 4.40。这说明 grounding 规则能显著约束模型的措辞，使其改用“我印象里你之前……”等试探性、不确定的方式提及旧记忆，而非断言为“你刚说”。两组场景均未出现编造共同经历的情况，说明该场景下模型的主要风险集中于“误引时机”而非“凭空捏造”。

Table 5. Anti-hallucination results (LLM-as-judge, 15 scenarios)
表 5. 反幻觉结果(LLM-as-judge, 15 场景)

条件	误引率	编造率	平均自然度
无 grounding 规则	0.400	0.0	3.33
有 grounding 规则	0.133	0.0	4.40

4.8. 局限性与讨论

本文评测基于合成构造的评测集，与真实对话数据仍存在分布差异，这是合成评测基准的固有局限。为缓解该偏差，生成过程遵循“近期性 - 相关性”“重要度 - 相关性”等真实行为假设，并对标注进行人工抽查；真实学习者的主观体验(如切换角色后的记忆连贯性感知)与长期记忆效果，仍需后续用户研究验证(见第 4 节)。此外，记忆抽取质量尚未纳入评测，相关评估留作后续工作。

5. 结论与未来工作

本文面向多角色 AI 学习伴侣场景，提出三级记忆可见性模型、加权检索排序与无关拒绝、以及记忆注入反幻觉约束。离线评测表明：该方法在召回与基线相当的前提下，实现无关查询的完全拒绝(6 条无关查询全部被拒绝)与跨角色私密记忆零泄漏；消融实验验证了时间衰减、重要度加权、模式加权与隐私分级各模块的明确贡献；grounding 规则将记忆误引率从 40.0%降至 13.3%。

未来工作包括：开展真实学习者的小规模用户研究，验证系统在实际使用中的记忆相关性体验与角

色切换时的连续性感知；扩大评测语料并引入记忆抽取质量评测；探索检索与生成端到端联合优化，以及将多角色记忆可见性模型推广到更广泛的教育与陪伴场景。

基金项目

浙江省大学生创新创业训练计划项目“基于大模型的风格化学习助手”(s202513291028)。

参考文献

- [1] Liu, D., Wu, Z., Song, D. and Huang, H. (2025) A Persona-Aware LLM-Enhanced Framework for Multi-Session Personalized Dialogue Generation. *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, July 2025, 103-123. <https://doi.org/10.18653/v1/2025.findings-acl.5>
- [2] Alonso, N., Figliolia, T., Ndirango, A., *et al.* (2024) Toward Conversational Agents with Context and Time Sensitive Long-Term Memory. <https://arxiv.org/abs/2406.00057>
- [3] Xia, M., Zhang, X., Dixit, S., *et al.* (2026) Memora: A Harmonic Memory Representation Balancing Abstraction and Specificity. <https://arxiv.org/abs/2602.03315>
- [4] Wang, Q., Fu, Y., Cao, Y., Wang, S., Tian, Z. and Ding, L. (2025) Recursively Summarizing Enables Long-Term Dialogue Memory in Large Language Models. *Neurocomputing*, **639**, Article 130193. <https://doi.org/10.1016/j.neucom.2025.130193>
- [5] 吴璇, 付涛. 检索增强生成技术研究综述[J]. 计算机工程与应用, 2025, 61(20): 19-35.
- [6] Wang, G., Nie, C., Wang, Z., Zhai, Y., Zhou, J., Haitao, S., *et al.* (2026) Dual-Perspective Cognitive Architecture for Long-Term Dialogue Memory. *Neurocomputing*, **700**, Article 134419. <https://doi.org/10.1016/j.neucom.2026.134419>
- [7] 张莉, 那蓉蓉, 张晓辉. 基于自适应问题分解的检索增强生成方法[J]. 中文信息学报, 2026, 40(2): 21-29+40.
- [8] 杨可颖, 冉洁茹, 刘瑞林, 等. 教育大模型在国产算力平台上的部署与实践——基于智谱清言大模型的个性化学习支持系统构建[J]. 计算机教育, 2025(12): 127-133.